VERIFICATION OF TRANSLATION

I, Ikuko Kamei, translator of 1-18-11, Minamikasahigashi, Kusatsu-shi, Shiga, Japan, hereby declare that I am conversant with the English and Japanese languages and am a competent translator thereof. I further declare that to the best of my knowledge and belief the following is a true and correct translation made by me of U.S. Patent Application No. 10/724,668 filed on December 16, 2003.

Date:
May 24, 2004

Ikuko Kamei
Ikuko Kamei

SPECIFICATIONS

Title of the Invention

Content Synchronization System and Method of Similar Web Pages

Background of the Invention

Field of the Invention

This invention relates to a content synchronization system and method of similar Web pages to display a basic Web page and similar Web pages similar to the basic Web page in an easily viewable manner.

Description of the Related Art

Presently, more than 36 million Web sites exist on the Internet. There are a great number of web sites consisting of more than 10,000 Web pages, which makes Web sites voluminous. To own a Web site is considered to be a matter of course for companies or universities. Accordingly, these Web sites can be classified into a similar field or a category of business, such as a portal site or a news site.

In order to compare pages of similar Web sites, a specific Web page is searched in a compared site or a page to be accessed is searched for each Web site and presented and then a comparison has to be made manually. For example, in order to compare how a news article is described for each site, a user has to open each site individually though a discrete Web Browser and to present related pages respectively and read them. Under such circumstances, it is difficult for a user to make a comparison on multiple sites.

Consequently, Dean et al. conducted a study to search a page related with a content of the Web page shown by a URL by means of giving the URL as search information. Their study makes use of not a content of the Web page itself but link information only or makes use of a description of customary information.

Taher. H et al. conducted a study in which a brother page of an input URL is searched by mans of a relationship between a parent page of the input URL and a child page quoted by the parent page by making use of a Web page link tree and ten nodes that are quoted most frequently are considered to be related pages.

However, since related information is obtained based on link information in either of these studies, a reliability to obtain similar pages is low and similar pages are just obtained and how to control a display mode of the similar pages is not mentioned.

Brief Summary of the Invention

The present claimed invention intends to make it possible for a user to read multiple similar sites in an easily viewable manner, when a page in a Web site is presented, by presenting a similar Web page in a different site based on a search keyword automatically obtained from the page simultaneously and automatically and also controlling a display mode of the similar page to be synchronous with a display mode of a basic Web page.

More concretely, as shown in Fig. 1, the present claimed invention relates to a content synchronization system of similar Web pages comprising a basic Web page specify portion that receives an identifier of a basic Web page as a Web page to be a basis of display and specifies the basic Web page, a compared Web site specify portion that receives an identifier of a compared Web site as a similar content of the basic Web page, a keyword extract portion that extracts a keyword in the basic Web page specified by the basic Web page finding portion, a similar Web page finding portion that specifies one or multiple similar Web pages that are similar to an entire part or a part of the basic Web page in the compared Web sites based on the keyword extracted by the keyword extract portion and a Web page display control portion that displays the similar Web pages specified by the similar Web page finding portion together with the basic Web page and synchronized with an operation of a user to the basic Web page. In the present claimed invention, a Web site to be a basis is called a basic Web site and a Web site to be compared is a compared Web site. In addition, a Web page in the basic Web site specified by a user is called the basic Web page and a Web page in a compared Web site similar to the basic Web page is called a similar Web page.

Brief Description of the Several Views of the Drawing

Fig. 1 is a block diagram showing a configuration of the invention in accordance with claim 1.

Fig. 2 is a pattern diagram of an equipment configuration of a content synchronization system of similar Web pages in accordance with an embodiment of the present claimed invention.

Fig. 3 is a pattern diagram of an equipment configuration showing an internal configuration of a terminal computer in accordance with the embodiment.

Fig. 4 is a pattern diagram of an equipment configuration showing an internal configuration of a center computer in accordance with the embodiment.

Fig. 5 is a functional block diagram of a content synchronization system of similar Web pages in accordance with the embodiment.

Fig. 6 is a tree structure explanatory diagram showing an example of a tree structure of Web pages in accordance with the embodiment.

Fig. 7 is a tree structure explanatory diagram showing another example of a tree structure of Web pages in accordance with the embodiment.

Fig. 8 is a table diagram showing a table structure of a Web configuration database in accordance with the embodiment.

Fig. 9 is a screen view showing an example of a screen transition of a similar Web page when a user clicks an anchor of a basic Web page.

Fig. 10 is a screen transition explanatory view showing a pattern of a screen transition of a similar Web page when a user makes an operation to scroll up or scroll down a basic Web page.

Fig. 11 is a screen view showing an example of a screen transition of a similar Web page when a user makes an operation to scroll up or scroll down a basic Web page.

Fig. 12 is a screen view showing a display mode of a difference Web page when a user clicks an icon of a similar portion.

Fig. 13 is a screen view showing an example of a screen transition of a similar Web page when a user makes an access to a previous or a next page again through a back or a forward function of a browser.

Fig. 14 is a screen view showing a display mode of a similar Web page when a user selects a word in a basic Web page.

Detailed Description of the Invention

An embodiment of the present claimed invention will be described in detail with referring to the accompanying drawings.

Fig. 2 is a diagram of an equipment configuration of a content synchronization system of similar Web pages in accordance with the embodiment. This system is so arranged that a terminal computer P1 (a terminal unit) to be used by a client is connected with a center computer P2 (a information processing unit) in a communicable manner and each computer is connected with the Internet.

The terminal computer P1 is a multi-purpose computer having, for example, a browsing function and comprises, as shown in Fig. 3, a CPU 101, an internal memory 102, an external memory unit 103 such as an HDD, a communication interface 104 such as a modem to connect with a communication network, a display 105 and an input means 106 such as a mouse or a keyboard.

The center computer P2 is a multi-purpose computer having, for example, a server function and comprises, as shown in Fig. 4, a CPU 201, an internal memory 202, an external memory unit 203 such as an HDD, a communication interface 204 such as a modem to connect with a communication network, a display 205 and an input means 206 such as a mouse or a keyboard. Each computer is not limited to a multi-purpose computer and it may be a computer for exclusive use or each computer is not physically separated and may be integrated into a single unit.

Explaining from a functional point of view, the center computer P2 is provided with a similar Web page search function and the terminal computer P1 is provided with an interface function. The similar Web page search function is a function to extract a keyword automatically out of Web pages in a basic Web site and to find a Web page similar to an entire part or a part of the basic Web site out of compared Web sites by making use of the above-mentioned keyword. The interface function is a function to extract a similar portion of an entire Web page or a part of a Web page and to present it to a user based on a behavior of the user.

More concretely, as shown in Fig. 5, predetermined programs are installed on each computer and the CPUs 101 and 201 and the peripheral equipment are operated together based on the programs so that this system fulfills its function as a basic Web page specify portion, a compared Web site specify portion, a keyword extract portion, a word frequency calculate portion, a Web page analyze portion, a Web configuration database, a similar Web page finding portion, a difference Web page finding portion, a Web page display control portion, a difference Web page display portion or the like.

Followings are descriptions of each portion and an explanation of an operation of this system.

I. Similar Web page search function

(1) Specify sites

A user designates a URL of a basic Web site and a URL of a compared Web site and selects a Web page that the user wants to browse from the basic Web site. At this time the basic Web site and the compared Web site have a similar content. An operation receive portion of the terminal computer P1 receives an operation by the user. Then the basic Web page specify portion and the compared Web site specify portion specify a site (or a page) based on the designated URLs and determine whether the information on the designated site (or the page) has already been registered in a Web configuration database. If the information is not registered, each of the specify portions obtains all real pages of each site from the Internet. (real page is described as real page information in Fig. 5) Next, an analysis result of each real page which will be explained in (2) and its result are registered in the Web configuration database. If the information is registered, an operation of (3) and below is conducted.

(2) Analysis and registration of Web pages

The Web page analyze portion makes a tree structure and analyzes paragraphs as shown in Fig. 6 and Fig. 7 by using structure tags of each Web page and identifies a title, a subtitle and a content.

The title and the subtitle are a word or a sentence surrounded by a tag in itself. Or it is often the case that each of the title and the subtitle is written in characters larger than those of other sentences in the Web page or that the characters of the title and the subtitle are emphasized. Then a word or a sentence enclosed by a <Font> tag or an <H> tag and ended with a noun or a proper noun as well is considered a title candidate or a subtitle candidate. A title is a word or a sentence that appears at the top of a Web page and locates at the shallowest and furthest left in the tree structure. A Subtitle is a candidate word or a candidate sentence other than the title. The title and the subtitle have a nested structure.

The word frequency calculate portion calculates a word frequency in a basic Web page, specifies a part of speech for each word by the use of Morphological analysis and obtains vectors for each word based on the word frequency of each word. More concretely, nouns are weighted based on a part of speech and each word vector is calculated with the word frequency multiplied by a word weight by the part of speech. The word weight assigned to each part of speech is, for example, 3.0 to a proper noun, 0.1 to a number, 0.1 to a numerical classifier, 1.0 to a general noun and 0.9 to other nouns.

The Web page configuration information as information on a Web page configuration analyzed by the Web page analyze portion or the word frequency calculate portion is stored in a Web configuration database. A table structure example of the Web configuration database is shown in Fig. 8.

(3) Extract of keywords

Next, the keyword extract portion extracts keywords from the Web configuration database. More concretely, a word contained in a title or a subtitle is extracted and the extracted word is considered a keyword for each of the title and the subtitle. At this time since the title and the subtitle have a hierarchical structure, the keyword is determined through a breadth-first search of the tree structure. In addition, if all words of nouns or proper nouns contained in a title or a subtitle are considered the keyword, a number of subject keywords might be too many. Accordingly, the word considered the keyword should have a word vector not less than a certain threshold α.

The title keywords **Ti** and the subtitle keywords **STxk** are considered subject keywords **inTitle**, where **i** is a number of title keywords, **x** is a number of subtitle keywords, and **k** is a number of keywords for a subtitle. The subject keywords **inTitle** is defined as

inTitle = (Ti, STij, ・・・, STxk)

Sentences other than the title or the subtitle are considered to show contents and then a content keyword is extracted. In order to obtain similarity for each part of the basic Web page, content keywords **inTexti**, **i∈(1, 2, ・・・, n)** are obtained from each paragraph of the basic Web page. The content keyword **inTexti** is a word whose word vectors is not less than a certain threshold α. The threshold α is equal to the threshold α of the word vector of the subject keyword. **i** shows a number of a paragraph. If a word is contained in a sentence showing its content and its word vector is not less than α is considered **Ci**, where **i=1, 2, ・・・, n**, the content keyword **inTexti** is defined as

inTexti = (C_o, C_1, ・・・, C_n)

The content keyword **inTexti** is ranked by the word vector out of the largest to the smallest.

The content keywords are stored in the Web configuration database.

(4) Search (specify) of similar page

Next, the similar Web page finding portion searches a similar Web page from the compared Web site by the use of the keyword searched and extracted from the basic Web page. Here dealt are a Web page entire part of which is similar to the basic Web page and a Web page whose part is similar to the basic Web page. The part of the Web page here means a paragraph of the Web page. The paragraph of the Web page is a node of a tree structure of the Web page using structure tags. In short in this embodiment, similarity search is conducted in a unit of a node of the tree structure of the Web page. A Web page whose entire part is similar to the basic Web page is a Web page that has the greatest number of similar nodes. A similar Web page similar to the basic Web page is determined from the compared Web site by the use of the subject keyword and the content keyword obtained in the former process of extracting keywords. Since that the subject keywords differ from the content keywords in meanings is experimentally proved, in this embodiment a subject keyword is searched from a title or a subtitle of a compared Web page of a compared Web site and a content keyword is searched from sentences showing a content in a compared Web page of a compared Web site. However, a Web page configuration without a subtitle differs significantly from one with a subtitle. As a result, a search is conducted differently for each case.

a) Web page having a subtitle

In this case, the Web page can be considered a structured Web page. As shown in Fig. 6, child nodes of a title and a subtitle are considered sentences indicating a nature of its content. Then in case that there is a subtitle in the compared Web page, the subtitle and its child node can be treated as a single entity and similar passages are searched as follows.

(1) A passage similar to a subject keyword is searched from a title and/or a subtitle of a compared Web page in the compared Web site. Since the title and the subtitle are within a nested structure, the tree structure is searched

through a breadth-first search. If the title and/or the subtitle is similar to the subject keyword, the content as its child node is also considered similar. As a result, no search is conducted for the child node of the title and/or the subtitle similar to the subject keyword. The similarity-degree is computed by the use of the Euclidian distance. In short, the title and/or the subtitle and its child node whose Euclidian distance from a subject keyword feature vector is the least are considered the similar paragraph.

(2) A passage similar to the content keyword is searched from content sentences. A passage similar to a content keyword is searched from sentences of nodes other than a child node of the node whose title and/or subtitle contains the subject keyword. In short, the node whose Euclidian distance from the content keyword feature vector is the least is considered the similar paragraph.

b) Web page without a subtitle

In this case, the Web page is considered a non-structured Web page. As shown in Fig. 7, a title is considered a root node, and other nodes are sentences indicating contents. In this case, all nodes are searched through the breadth-first search and a node similar to the content keyword is searched. In short, the node whose Euclidian distance from a content keyword feature vector is the least is considered the similar paragraph.

Paragraphs similar to the basic Web page are found for each compared Web page in the compared Web site. A Web page having the greatest number of similar paragraphs is a similar Web page candidate. If multiple Web pages are candidates to become the similar Web page, the one with the shallowest node and farthest left node in the link tree of the compared Web site is selected as the similar Web page.

(5) Obtain information on difference between the basic Web page and the similar Web page

All the content contained in the basic Web page is not contained in the similar Web page. There are some cases that other page in the compared Web site has some information that is contained in the basic Web page and that is not contained in the similar Web page.

Then in this embodiment a Web page having difference information between the basic Web page and the similar Web page is presented in other window. In the former processing, in the Web page of the compared Web site, a similar paragraph similar to the basic Web page is searched and specified for every paragraph.

The difference Web page finding portion finds a paragraph whose similarity-degree of the subtitle keyword **STxj** or the content keyword **inTexti** contained in the paragraph of the basic Web page that does not have a similar paragraph in the similar Web page is the highest from Web pages in the compared Web site other than the similar Web page. The Web page having this paragraph becomes a difference Web page having difference information between the basic Web page and the similar Web page. If there are multiple difference Web page candidates, the one with the shallowest node and farthest left node in the link tree of the compared Web site is selected as the difference Web page.

II. Interface function

An interface function is a function to present a similar Web page together with the basic Web page and synchronized with an operation of a user such as clicking, scrolling, navigating forward and backward and the Web page display control portion arranged on the terminal computer P1 serves as its function. The user gets a view of this portion.

The interface function will be explained concretely.

(1) Presentation of similar Web page when a user clicks

As an example of a display is shown in Fig. 9, when the user clicks an anchor of the basic Web page, a linked page becomes a new basic Web page. Then a keyword is extracted from the new basic Web page, a similar Web page is found from the compared Web site and presented synchronously with the basic Web page. At this time a Web page that has difference information between the basic Web page and the similar Web page is presented in a form of an icon.

(2) Presentation of similar part of similar Web page when a user scrolls

There are a lot of Web pages wherein a length of a page is long. In this case, a user scrolls up or down a window in order to browse this Web page. Then in this embodiment when a user scrolls up or down the basic Web page, a paragraph in a similar Web page that is similar to a paragraph in the basic Web page is automatically scrolled up or down and presented to the user. A pattern diagram is shown in Fig. 10 and an example of a display is shown in Fig.

11. In case of no similar paragraph in the similar Web page as shown in Fig. 12, if a user clicks an icon of the similar part, the difference Web page display portion displays the difference Web page having difference information between the basic Web page and the similar Web page on a different window.

(3) Presentation of similar Web page when a user navigates backward or forward

When a user browses a previous or a next page again by the use of back or forward function of a browser, the basic Web page and the similar Web page are presented synchronously as an example of a display is shown in Fig. 13.

(4) Presentation of similar Web page when selecting a word in the basic Web page

In this embodiment, a user browses two different Web pages at once. In this case, however, it is conceived that similar information is difficult to obtain at a glance. As a result, as an example of a display is shown in Fig. 14, a word selected by a user in the basic Web page is presented in a mode different from other words such that the selected word is highlighted in the similar Web page so that the user can obtain similar information viscerally.

III. Summary

As mentioned above, the system in accordance with this embodiment extracts a keyword from a basic Web page in a basic Web site specified by a user, automatically finds a similar Web page from compared Web sites by the use of the keyword and presents it simultaneously. The keyword comprises a subject keyword and a content keyword and the subject keyword is used for searching a title and/or a subtitle and the content keyword is used for searching contents. The similar Web page is found by the use of a tree structure of a Web page configuration. By using this system a user can browse a similar Web page in compared Web sites with ease just by browsing Web pages one by one in the basic Web site sequentially.

The present claimed invention is not limited to the embodiment. There may be various modifications without departing from a spirit of this invention, for example, multiple similar Web pages similar to the basic Web page may be presented simultaneously and synchronously.